DATA MINING V

DATA MINING, TEXT MINING AND THEIR BUSINESS APPLICATIONS

WIT PRESS

WIT Press publishes leading books in Science and Technology.
Visit our website for the current list of titles.
www.witpress.com

WIT eLibrary

Home of the Transactions of the Wessex Institute, the WIT electronic-library provides the international scientific community with immediate and permanent access to individual papers presented at WIT conferences.
Visit the WIT eLibrary at www.witpress.com

FIFTH INTERNATIONAL CONFERENCE ON DATA MINING

DATA MINING V

Conference Chairmen

A. Zanasi
TEMIS Text Mining Solutions, Italy
N.F.F. Ebecken
Federal University of Rio de Janeiro, Brazil
C.A. Brebbia
Wessex Institute of Technology, UK

International Scientific Advisory Committee

Organised by
Wessex Institute of Technology, UK

DATA MINING V

DATA MINING, TEXT MINING AND THEIR BUSINESS APPLICATIONS

Editors

A. Zanasi
TEMIS Text Mining Solutions, Italy

N.F.F. Ebecken
Federal University of Rio de Janeiro, Brazil

C.A. Brebbia
Wessex Institute of Technology, UK

WITPRESS Southampton, Boston

A. Zanasi
TEMIS Text Mining Solutions, Italy

N.F.F. Ebecken
Federal University of Rio de Janeiro, Brazil

C.A. Brebbia
Wessex Institute of Technology, UK

Published by

WIT Press
Ashurst Lodge, Ashurst, Southampton, SO40 7AA, UK
Tel: 44 (0) 238 029 3223; Fax: 44 (0) 238 029 2853
E-Mail: witpress@witpress.com
http://www.witpress.com

For USA, Canada and Mexico

WIT Press
25 Bridge Street, Billerica, MA 01821, USA
Tel: 978 667 5841; Fax: 978 667 7582
E-Mail: infousa@witpress.com
http://www.witpress.com

British Library Cataloguing-in-Publication Data

A Catalogue record for this book is available
from the British Library

ISBN: 1-85312-722-9
ISSN: 1470-6326

The texts of the papers in this volume were set individually by the authors or under their supervision. Only minor corrections to the text may have been carried out by the publisher.

Printed in Great Britain by Athenaeum Press Ltd. Gateshead.

Preface

The Data Mining Conference in 2004 was again held in Europe, in Malaga, Spain, after previously taken place in Rio de Janeiro, Brazil, which hosted the 4th Conference.

This book contains the papers presented at the 5th Data Mining Conference. The Conference is becoming a traditional meeting for the practitioners of data mining. As data mining is evolving, so is the conference, although it remains faithful to its original formula of providing a meeting point for practitioners of data mining from academia and industry, as can be seen in the contributions published in this book.

Nearly 25 countries from every continent are represented in this book. The ISAC-International Scientific Advisory Committee includes high level scientists and professionals representing academia (from North and South America, Europe and Asia), industries (IBM, Microsoft, Siemens, SAS, SPSS, Insightful, Megaputer, Temis) and analysts (Gartner Group and META Group).

Data Mining is evolving, but in what sense and in what direction? The definition of data mining as "the process of extracting previously unknown, valid and actionable information from large databases and then using it to make crucial business decisions"[1] is still valid in the sense that many other sectors, such as artificial intelligence, applied mathematics, control theory and information technology still maintain an important role in providing techniques and subjects of research interesting to data mining.

It is also important that the concept of "large databases" be understood in a wider meaning by taking into account the, so called, "internet sources." Until recently mining has been, in fact, relegated to well-structured internal information. Mining the public web has been an exercise in frustration, but it is now possible to apply the same mining tools and techniques to the explosion of data found on billions of web pages, newsgroups, emails, chatlines, forums, blogs and message boards[2].

---

[1] Cabena, Hadjinian, Stadler, Verhees, Zanasi-*Discovering Data Mining*-1998-Prentice Hall

[2] A.Zanasi-Email, chatlines, newsgroups: a continuos opinion surveys source thanks to text mining application – 2003 –in *Excellence 2003 in Int'l Research* - Ed.Esomar

In this sense, thanks to the explosion of the Internet phenomenon and of new sources, data mining applied to text, ie. text mining, has acquired much more importance.

Researchers coming from other disciplines of science and research (eg. information retrieval, semantic web, linguistics, knowledge management…) are starting to show interest in *unstructured data mining* (ie text mining) results for their studies and applications.

This is the reason why it was decided to add a clear reference to text mining in the title of this conference, which now appears as "Data Mining, Text Mining and its Business Applications", to be retained also for the next conference (Data Mining 2005, to be held on 25 May 2005 in Skiathos, Greece).

The reference to "Business Applications" is there to stress that contributions regarding solutions to the problems that the real world pose to us are particularly welcome.

In recent years, businesses operating in diverse areas have shown an interest in data mining. Three of them appear to be the most promising, i.e.

National Security. Reading magazines, listening to television all over the world, no matter where one lives, there is no doubt that the most pressing problems are those linked to terrorism and to national security.
It is not a coincidence that more books are being published on data, text mining and business intelligence techniques applied to intelligence or national or homeland security.

In the USA fierce discussions about the opportunity of fighting terrorism using data and text mining capabilities, organized under DARPA (eg Information Awareness Office and Terrorist Information Awareness programs), are still firing the political debate, fearing the risk that this means to data privacy and to democracy itself. The CIA launched In-Q-Tel, a non profit corporation designed to fuel private research, development and application in, amongst others, data and text mining technologies[3]. It is to be expected that in the next few years this business area will grow quickly.

Competitive Intelligence (CI). As one of the developing software-based new analysis techniques, data and text mining provide a powerful way to uncover changing business issues and discover emerging trends that can drive business. CI capabilities are also changing as professionals pick up the tools that can help them from being reporters of facts, to predictors of market changes.

Combining the precision of analysis with the serendipity of discovery, text mining helps detect the complex signals of emerging trends over billions of pieces of information. Together with visualization tools, text mining provides a unique window to the world[4].

---

[3] John Gannon-*The Strategic Use of Open-Source Information* –in Intelligence and the National Security Strategist: Enduring Issues and Challenges-Sherman Kent Center for Intelligence Studies National War College National Defense University (2004)

[4] Dennis Cahill – Visualizing Emerging Intelligence through Text Mining – in Competitive Intelligence Magazine, May-June 2004

Customer Relationship Management (CRM). Many organizations already apply data mining to internal customer information. Using large Customer Relationship Management (CRM) tools and other internal databases, companies identify in a detailed way how customers purchase (or don't purchase) products. In the high-volume consumer transactions of ecommerce business, analyzing information through data mining can yield intelligence that can affect revenue or profit. The opportunity of analyzing emails and chatlines, thanks to text mining, allow for the detection of customers' opinions and trends to boost sales. It is not a coincidence that a section of this book is dedicated to CRM.

For a discipline that is central to research and business, and which is of interests not only among software developers, but also among large companies and government departments and agencies, it is important to define its market size. IDC estimate, for only the data mining market size in 2005[5], $ 1.5 billion, and OVUM estimate for data and text mining and the related technologies, grouped under the term "knowledge management"[6], $ 11.98 billion. These are large sums, and seeing the trend of growing government expenses in the national security market, where data and text mining are of fundamental importance, this sum will continue to grow.

The success of the Conference and this resulting book would not have been possible without the generous help of the members of the I.S.A.C. Their efforts ensured the quality of the papers published in this volume.

Alessandro ZANASI

Malaga, September 2004

[5] IDC-Information Access Tools : Market Forecast and Analysis
[6] SIIC Journal: http://www.siic.org.hk/journal/935.582.107.PDF

Contents

SECTION 3: TECHNIQUES

Section 3a: Clustering

Section 3b: Categorization

Section 3c: Link analysis

Section 3d: Data preparation

SECTION 4: APPLICATIONS IN BUSINESS, INDUSTRY AND GOVERNMENT

SECTION 5: CUSTOMER RELATIONSHIP MANAGEMENT